BIG DATA & SOCIETY

Original Research Article

Big Data & Society
July–December 2015: 1–11
© The Author(s) 2015
DOI: 10.1177/2053951715611145
bds.sagepub.com
SAGE

Known or knowing publics? Social media data mining and the question of public agency

Helen Kennedy[1] and Giles Moss[2]

Abstract

New methods to analyse social media data provide a powerful way to know publics and capture what they say and do. At the same time, access to these methods is uneven, with corporations and governments tending to have best access to relevant data and analytics tools. Critics raise a number of concerns about the implications dominant uses of data mining and analytics may have for the public: they result in less privacy, more surveillance and social discrimination, and they provide new ways of controlling how publics come to be represented and so understood. In this paper, we consider if a different relationship between the public and data mining might be established, one in which publics might be said to have greater agency and reflexivity vis-à-vis data power. Drawing on growing calls for alternative data regimes and practices, we argue that to enable this different relationship, data mining and analytics need to be democratised in three ways: they should be subject to greater public supervision and regulation, available and accessible to all, and used to create not simply known but reflexive, active and knowing publics. We therefore imagine conditions in which data mining is not just used as a way to know publics, but can become a means for publics to know themselves.

Keywords

Social media data mining, data mining, publics, agency, knowing publics, calculated publics

Introduction

Social media data mining is on the rise. The increasing availability of data on users and their online behaviour, the decreasing cost of collecting, storing and processing data, and the exponential expansion of social media platforms from which much of this data is taken mean that – at least in theory – an increasingly diverse range of actors can mine social data. This process can involve simply counting the likes and shares of social media content, or more advanced analysis of its strength, sentiment, passion, reach and other quantifiable characteristics (mentions, users, sources, hashtags). The metadata that sits behind social media content is also widely mined, and considered by some to be more valuable than the content itself. Such metadata includes: who is speaking and sharing, where they are located, to whom they are linked, how influential and active they are, what their previous activity patterns look like and what this suggests about their likely preferences and future activities. Social media data mining is undertaken by the major platforms themselves (like Facebook and Twitter), by intermediary commercial companies (such as Sysomos, Radian6, Brandwatch), or with tools which are free to all comers; some easy-to-use (for example Social Mention), and others more complex (such as NodeXL).

Methods for analysing social media data promise powerful new ways of knowing publics and capturing what they say and do. And yet access to these methods is uneven, with large corporations and governments tending to have the best access to data and analytics tools. Critics warn of a number of

[1] Department of Sociological Studies, University of Sheffield, Sheffield, UK
[2] School of Media and Communication, University of Leeds, Leeds, UK

Corresponding author:
Helen Kennedy, Department of Sociological Studies, University of Sheffield, Sheffield S103TN, UK.
Email: h.kennedy@sheffield.ac.uk

troubling consequences for publics that result from the rise and spread of data mining: less privacy, more surveillance and social discrimination, and a new means of controlling how publics come to be represented and so understood. Meanwhile, the tools and systems that generate knowledge from social media data are typically opaque and are rarely open to public scrutiny and supervision. We argue that these various enactments and characteristics of data mining are constitutive of a new form of data power.

In the light of data power, this article considers whether a more positive relationship between social media data mining and public life might be established, one in which publics can be said to have greater agency and reflexivity. We argue that, in order for this to happen, there is a need to democratise data power in three main ways. First, to address concerns about the potential negative effects of data mining on the public, it needs to be subject to greater public supervision and regulation. Secondly, to address the danger of new, data-driven digital divides emerging, the technologies of data mining (which include software and expertise as well as data themselves) must be available and accessible to the public so they can be used in varied ways. Thirdly, given the contribution that data mining increasingly makes to how publics and public issues are represented, data mining could be used in ways that enable members of the public to understand each other, reflect on matters of shared concern, and decide how to act collectively as publics, thereby allowing publics to constitute themselves as more reflexive and active agents.

Together, these three ways of democratising data mining (subjecting it to greater public supervision and control, ensuring it is available and accessible to the public to use, and using it in a way that enables the production of more reflexive and active publics) address concerns expressed about data mining and point us towards ways in which more *knowing* publics (rather than just *known* publics) might surface through data mining. They may, therefore, produce conditions in which the public can act with greater agency in relation to data mining, in the sense that Couldry (2014: 891) has defined this term: 'not brute acts (of clicking on this button, pressing like to this post)' but rather 'the longer processes of action based on reflection, giving an account of what one has done, even more basically, making sense of the world *so as* to act within it'. Baack (2015) argues that thinking about agency is fundamental to challenging the structures of data power and yet questions about agency have been 'obscured by unnecessarily generalised readings' (Couldry and Powell, 2014: 1) of the supposed power of technological assemblages like data mining. For this reason, Couldry and Powell (and others) call for more attention to agency than theories of algorithmic power, or data power, have thus far made possible. This paper represents one such endeavour.

Our paper is an imagining of the conditions that are required to democratise data mining, grounded in examples of the three strategies we discuss. As such, our approach is normative, and is in line with those grounded critical theories that seek both to analyse problems with current social practices and articulate what might be valuable about them (Young, 2002: 11–12). We are aware that despite growing calls to think about and do data mining differently, efforts to democratise data mining are far from being realised in practice. Nonetheless, we feel it is important to do the work of imagining what could be with regard to the relationship between data mining and publics. Given that, as van Dijck and Poell (2013) assert, all kinds of actors (in education, health, politics, arts, entertainment, policing, activism) are increasingly required to act within what they define as 'social media logic' (which includes social media data mining), imagining how such practices might be more democratic seems like a vital undertaking. In doing this, we revisit the question that Andrew Feenberg asked in his preface to *Transforming Technology*: 'must human beings submit to the harsh logic of machinery, or can technology be fundamentally redesigned to better serve its creators?' (2002: v).

Although more and more data are mined from an ever broader range of sources, we focus on social media data mining here for three reasons. First, because a wide range of public actors are *technically* able to engage in it, as the open APIs (Application Programme Interfaces) of social media platforms make it possible for non-corporate actors to analyse at least some public social media data. Second, while social media have been viewed as sites of 'interactivity' among publics, social media data mining is categorically not interactive: it takes data from social media and analyses them, and publics are not able to intervene or interact in this process. For this reason, the politics of social media data mining need critical attention. Third, social media have been viewed as crucial sites where publics emerge. Described first as 'networked publics' assembling in and structured by social media platforms (Boyd, 2010) and then as algorithmically generated 'calculated publics' (Gillespie, 2014), in this paper we consider whether it is possible for the publics that take shape through data mining to be characterised by agency and so be understood as neither known nor calculated publics, but rather as knowing publics. We proceed to elaborate on some of the main criticisms that have been levelled at social media data mining's production of known publics, in order then to consider the conditions under which more knowing publics might emerge.

Known publics

The analysis of social media data is seen as a powerful new way of knowing publics and capturing what they say and do. But social media data mining and analytics tend to be dominated by corporate and government elites, who generally have the best access to data and analytics tools. Critics warn of a number of problems for the public that may stem from the growing data power of these groups: it is likely to result in less privacy and more surveillance, increased social discrimination and 'deep personalisation' (Couldry and Turow, 2014: 1712), and it provides the already-powerful with control over how networked publics come to be represented and understood.

The most common concern that arises from corporate and governmental uses of social media data mining to know publics is that it results in less privacy and more surveillance. In 2010, Facebook CEO Mark Zuckerberg announced that in the age of social media, privacy is no longer 'a social norm' (Johnson, 2010). At the same time, a number of commentators have contested this view, such as Boyd (2014), whose extensive ethnographic research into teen social media attitudes leads her to argue that privacy still matters to young people. Contrary to Zuckerberg's assertion, privacy still is a social norm, she and others claim. Of course, it is in the interests of social media companies who make money by selling the content that users share on social media platforms to tell us that we no longer care about privacy – indeed, such strategies play a role in shaping how we think. Yet despite the efforts of Zuckerberg and others to dismiss the significance of privacy in social media environments, the concept retains traction. Examples of the invasion of social media privacy by corporations are greeted by public concern, and academic researchers seek to understand what they call the 'privacy paradox', or the fact that social media users' sharing practices appear to contradict their expressed privacy concerns. Some authors conclude that for users, there is a distinction between social privacy (controlling which people within their networks get access to their information) and institutional privacy (the mining of personal information by social media platforms and other commercial companies) (Raynes-Goldie, 2010; Young and Quan-Haase, 2013).

Alongside concerns about personal privacy invasion, social media data mining is seen as having increased surveillance. Trottier (2012) argues that social media data mining opens up access to aspects of life once intimate and guarded and, as such, is a new form of surveillance (see also Fuchs, 2014). Andrejevic provides a convincing array of examples to argue that there are various ways in which our actions are subjected to surveillant scrutiny, including forms of social media data mining like sentiment analysis and opinion mining, but also other activities like body language analysis, neuromarketing and drone technology (2013). Social media platforms like Facebook are ideally situated for ubiquitous surveillance, he argues with Gates (Andrejevic and Gates, 2014), as they have the infrastructure in place that makes it possible for them to serve as surveillance systems. In a recent online list of the 10 largest databases in the world, a number of social media platforms figure (Anonyzious, 2012, cited in Andrejevic and Gates, 2014: 189). Thus corporate and governmental uses of social media data mining can be seen as contemporary forms of surveillance, and platforms are simultaneously actors in surveillant practices, infrastructures that enable surveillance, and databases which house the datasets that are surveilled.

With all of this mined data comes the possibility of discriminating amongst members of the public, and a number of scholars have pointed to the various domains in which this discrimination is taking place. Turow (2012) highlights how the digital advertising industries use data to discriminate as, through data analytics processes, 'individual profiles' are turned into 'individual evaluations' (p. 6). Based on behavioural and other forms of tracking, individuals' marketing value is calculated and each individual is categorised as target or waste. Like Turow, others have highlighted how the discriminatory potential of data mining is captured in the interests of capital. Hearn (2010) argues that data mining's ability to identify valuable sentiments monetises feeling and intimacy and represents yet another capitalist mechanism of value extraction, and Andrejevic (2011) writes about the role sentiment analysis plays in the prediction and control of affect. Elsewhere, Beer and Burrows (2013) point to the ways in which data-based discrimination operates in the production of culture. Writing about music consumption technologies and their generation of archivable data about listening habits, they suggest that such data do not only constitute listening practices, but also 'feed into the production of large-scale national geodemographic systems that in turn provide postcode-level analysis of people's tastes and preferences' (2013: 59). As such, data constitutes much more than culture, serving also to shape regimes of governance and control (see also Barocas and Selbst, 2014).

Elsewhere, commentators have raised concerns about the democratic implications of the 'deep personalisation' that result from the discriminatory practices discussed above (Couldry and Turow, 2014: 1712; Kant, 2014; Pariser, 2011). Pariser (2011: 24) warns that the resulting 'filter bubble' of media content tailored to individuals is a 'centrifugal force' that separates members of the public from one other and weakens democracy. Couldry and Turow (2014: 1716–1718) describe how personalisation might shape the content media organisations produce, as they adapt news and other content based on the knowledge they have about audiences. Deep personalisation of media, they argue, threatens to undermine the 'shared reference points' necessary for democracy and which 'enable us to recognize one another as members of a common social and political space' (Couldry and Turow, 2014: 1719).

At the same time, some commentators have argued that data-mining techniques are used by dominant groups as a powerful new way of controlling how publics are represented and so understood. In this respect, if personalisation is a centrifugal force (Pariser, 2011: 24), data mining can act as a centripetal force in representing what broader publics are saying and doing. While much knowledge generated through social media data is not shared with the public, representations of the public based on social media data now circulate in the public sphere and shape how publics are viewed and how they see themselves (Anstead and O'Loughlin, 2015; Gillespie, 2014). Discussing the UK general election in 2010, Anstead and O'Loughlin (2015: 215) note that 'the use of social media as a tool to understand and illustrate public opinion is starting to enter into mainstream media discourse', through 'more complex semantic polling techniques' as well as 'electronic vox pops and commentary on trending topics'. The public, they argue, 'reflect on' 'interpret and talk about these forms of public opinion in their everyday lives' (Anstead and O'Loughlin, 2015: 216). Similarly, Gillespie (2014) discusses how representations of the public generated through data – what he terms 'calculated publics' – are becoming increasingly prominent ways of thinking about publics. He gives the example of Twitter's Trends algorithm, which claims to represent what different geographical publics are discussing on Twitter at particular moments in time (Gillespie, 2012). Gillespie (2014) argues that these representations matter because they play a role in shaping and constituting what we take publics to be. He therefore urges us to ask, 'how do these technologies, now not just technologies of evaluation but of representation, help to constitute and codify the publics they claim to measure, publics that would not otherwise exist except that the algorithm called them into existence?' (2014: 189). The publics generated through data are contestable: they do not simply mirror publics 'out there', but rather are constructed in particular and partial ways. Given this, Gillespie (2014: 189) suggests there is the possibility of a 'friction between the "networked publics" forged by users and the "calculated publics" offered by algorithms'.

We argue below that the representational work data mining can perform in representing broader publics – especially in view of the personalisation and fragmentation of digital spaces – could be a valuable resource in providing 'shared reference points' that Couldry and Turow (2014) argue are crucial for democracy. The important question is to what extent data mining is used to promote public reflection, understanding and debate, and whether the representations it generates are contestable. Gillespie (2014: 191) suggests that with each public medium (now algorithms, previously, for example, newspapers), we 'turn over the provision of knowledge to others' and so are 'left vulnerable to their choices, methods and subjectivities'. Sometimes this can be positive, he argues, as it provides us with knowledge filtered through editorial expertise, as in the case of newspapers. But sometimes it is not, because filtering and selection procedures are unavailable to us, or because they have undesirable social and political consequences. Algorithms are 'socially constructed and institutionally managed mechanisms for assuring public acumen: a new knowledge logic', he writes (Gillespie, 2014, 192). However, the criteria by which calculated publics are generated are typically unknown and not clearly explained (Gillespie, 2014).

The concentration of data power in elite hands and the related unequal access to data and analytics tools can make it difficult for publics to contest or challenge the ways in which data mining plays a role in representing publics. Indeed, another concern expressed in relation to the rise of data mining, one of Boyd and Crawford's (2012) six 'provocations for Big Data,' is that limited access to Big Data is creating new digital divides. Boyd and Crawford (2012) point out that while much of the enthusiasm surrounding Big Data comes from a belief that they are easy and straightforward to access, this is categorically not the case. So our statement in the introduction, that it is *technically* possible for a range of actors to engage in social media data mining because of platforms' open APIs, requires the important caveat that there are a number of challenges to realising this possibility. At present, elite commercial companies like Google, Facebook and Amazon have the best access to data, as well as the best tools and methods to make sense of it (Williamson, 2014). Some companies restrict access to data entirely, others sell access for a high fee, some offer small data sets to university-based researchers. Thus those with money or inside a company have differential access to social media data from those without financial resources or operating outside the major companies.

These data inequalities relate not only to data and analytics tools, but also to the expertise needed to use and make sense of them. Manovich states that there are three ways of relating to Big Data: there are 'those who create data (both consciously and by leaving digital footprints), those who have the means to collect it, and those who have expertise to analyze it' (2011: 10). Boyd and Crawford (2012) also point out that who is deemed to have expertise determines both who controls data-mining processes and the 'knowledge' about publics that results, knowledge which in turn produces publics, as we suggest above. 'Wrangling APIs, scraping, and analyzing big swathes of data is a skill set generally restricted to those with a computational background' (2012: 674), they argue. So access to data-mining expertise, as well as to data themselves, is uneven, and this produces new digital divides in relation to access to data, tools, skills and expertise (Boyd and Crawford, 2012).

The criticisms of social media data mining outlined in this section point to problems with current dominant uses of data mining. Critics argue that the imperative of government and corporate elites to know publics in particular ways through social media and other forms of data mining leads to the erosion of personal privacy and a parallel growth in practices of surveillance, discrimination and elite control over the ways in which publics are represented and known. These are all valid concerns, but they leave us with the question of whether it has to be so, or whether data mining may be reimagined in ways that allow it to contribute more positively to public life. In the next section, we describe alternative ways of thinking and doing data mining that seek to democratise data power and which suggest that a different relationship between the public and data mining may be possible. Our intervention, then, is to suggest ways of moving beyond critique. We characterise the battle over data power as being about a move from 'known publics', who are subject to the data-mining practices of powerful groups, to 'knowing publics', who are more active and reflexive agents.

Knowing publics

We have described how data mining is used by corporations and governments to know publics and how these practices raise concerns about privacy, surveillance, social discrimination and control over how publics are represented and understood. These problems are exacerbated by the emergence of new digital divides and inequalities around data. The critiques of data mining we have discussed above do the important job of highlighting some of the troubling consequences of current forms of social media data mining. However, they only take us so far: they do not tell us whether alternative data practices and arrangements are possible and, if so, what they should look like. In this section, we take critiques of data mining as a starting point and ask: 'given these problems, what then?' Looking towards alternative ways of thinking about and doing data mining that are emerging, we consider how data mining may be reimagined in ways that allow it to contribute more positively to public life.

Following the concerns expressed about social media data mining, there have been growing calls to think about and do data mining differently and so to democratise data. We highlight three main aspects to this politics of data mining, all of which seek to increase the power of the public vis-à-vis current data-mining regimes:

1. Data-mining practices should be subject to greater public supervision and regulation.
2. Data mining (data, tools, and expertise) should be accessible for all to use.
3. Data-mining practices should be used in ways that help to make more reflexive and active publics.

The first two of these are already the subject of widespread discussion, but the third is less widely discussed in relation to data mining and analytics. Our contribution to this debate is to argue that all of these three ways of democratising data mining are necessary to address the problems of data power, because together, they provide the means by which publics may be empowered through data. Of course, these are not straightforward solutions and are far from being implemented in practice, but identifying and outlining them is a necessary part of our project of imagining alternative and more democratic forms of social media data mining.

Perhaps the most widely discussed way in which the public can have more control over data mining is by making data-mining practices more transparent. Concerns about the negative social consequences of data mining in terms of privacy, surveillance, social discrimination, personalisation and control over how the public are represented and exacerbated by the black-boxing of data-mining processes: it is difficult to evaluate data-mining practices because code, algorithms and methodologies are often proprietary and we do not always know how they work. Gillespie (2014) highlights this problem when he talks about the opacity of Twitter's Trend algorithm and digital reputation measurement platforms like Klout. Given this, commentators argue that data-mining techniques need to be more transparent (Anstead and O'Loughlin, 2012; Couldry and Powell, 2014). Proponents of this position argue that making data-mining algorithms and processes public in this way would help to facilitate public understanding, scrutiny and debate about the political effects of data mining, and allow the public and groups acting on the public's behalf to examine and contest data-mining practices.

Some efforts to communicate transparently about data-mining practice are starting to emerge, although not necessarily on the major social media platforms. The UK broadcaster Channel 4 attempted to communicate its uses of viewer data to viewers through the production of a video featuring one of its comic talents, Alan Carr, in which the comedian describes which data the broadcaster asks its viewers to share voluntarily, the uses to which they are put, and the benefits to viewers (http://www.channel4.com/4viewers/). Similarly, the dating website OK Cupid attempted to explain its algorithmic matching processes in an animated video called 'the math of online dating' (https://www.youtube.com/watch?v=m9PiPlRuy6E) and the site also regularly unpacks the platform's own data on its blog (http://blog.okcupid.com/).

However, some commentators argue that making data mining transparent is not enough. Requiring companies to show their algorithms does not mean they will or are required to revise problematic practices, nor does it necessarily lead to greater public understanding, given the levels of expertise required to make sense of the technical operations of data-mining processes. What is needed, then, is not transparency, but accountability, some argue (Diakopoulos, 2014; MacKinnon, 2014; Pasquale, 2015; Sandvig et al., 2014). As Couldry and Powell (2014: 4) note, transparency goes some way towards addressing the problems of data mining's opacity and black-boxing, but it still 'fails to address accountability and reflexivity'. In other words, transparent companies are not necessarily accountable. Here, accountability might be understood in the terms in which Giddens (1984: 30) defines it: 'to be "accountable" for one's activities is both to explicate the reasons for them and supply normative grounds whereby they can be "justified"'. Drawing on Giddens' definition, McQuail (2003: 15) writes that 'we can view accountability as the entire process (within a communication relationship) of making claims based on expectations and appeals to norms, the response of the other party (rejecting claims or explaining actions), and any ensuing procedures for reconciling the two'. In the context of democratising data mining, accountability would therefore mean requiring data-mining companies not just to *show* the public what they are doing, but to *tell* publics what they are doing, why, and with what effect. Such accountability makes it possible to audit firms in a way that transparency does not, and it is for these reasons that proponents cited here argue for 'algorithmic accountability' rather than transparency.

Other commentators stress the need to regulate the uses to which data mining is put in order to prevent harm to the public (Barocas et al., 2013; Zarsky, 2004). Self regulation by private companies is not likely to be sufficient. To secure the public interest, as Freedman (2012) argues in relation to the Internet and the media industries more generally, regulation requires the intervention of public authorities. But any government regulation of data mining must itself be democratic and accountable to the public, not least because of the use of data mining and analytics by governments themselves. Government regulation imposed from the top down, without input from the public, would not guarantee that it serves the public and addresses public concerns. The public needs to be involved in determining how data mining will be regulated, if regulation is to be legitimate and enjoy public support. Examples which try to enact this proposal on a small scale include the EU's Hack4Participation initiatives: these hackathons explore how to get EU citizens more involved in EU policy-making and how to enable the better analysis of policy-making processes, and they sometimes result in the development of data-related policy. For example, policy relating to Net Neutrality was developed at a German policy hackathon and the Icelandic Modern Media Initiative operates in this way, adopting a strategy not of lobbying but of writing media policies which are relevant to the current digital age (for example in relation to privacy) (Hintz, 2014).

The second way to democratise data mining relates to public access. Given concerns about the emergence of a new digital divide around data, some commentators argue that if data mining is to serve the objectives of the public, it needs to be accessible to the public, not just major corporations, governments and security agencies. One proposed solution to problems of access which is relevant to our focus here is open data. According to Bates (2013), open initiatives like Open Government Data, Open Access and Free and Open-Source Software can be understood as efforts to reverse the trend towards the private ownership of and differential access to data that results in the kinds of known publics discussed above. Open data groups lobby for access to and the ability to re-use datasets, often focusing on those produced by public institutions. They insist on access and re-use for everyone, 'free of charge, and without discrimination' (Bates, 2013: np). Such groups see the opening up of public datasets as a form of democratisation of data, allowing the access to data that Boyd and Crawford argue is ominously absent from the data delirium (van Zoonen, 2014). But returning to the arguments of Manovich and Boyd and Crawford, open data advocate Rae (2014) argues that although the release of open datasets can be for the public good, it needs to be accompanied by skilled analysis and, importantly, by the right answers to these three questions: opened by whom, open to whom, and open to what? Open access to data is only one step in overcoming the danger of data-driven digital divides and in making data more accessible to publics. Alone, it is not an unproblematic solution to the problems of data mining discussed above. Indeed, as Bates and others (for example Gurstein, 2011) point out, there are many ways in which open data may serve to empower the technologically-elite and already-empowered – Bates' own study into open government data confirmed that governments usually release open data under conditions that allow them to control information flows.

It should be noted that these discussions are about the democratisation of access to data produced by public institutions, not to data mining, nor to social media data. Couldry and others addressed the issue of making data mining and analytics public – or social, in their terms – on a project called Storycircle (http://storycircle.co.uk/), which aimed to develop understanding of how digital resources can support individual and group agency in a specific social context (Couldry, 2014). One such resource is analytics. Couldry and collaborators developed the notion of 'social analytics' to describe their research into social actors' uses of analytics, what Couldry describes as 'alternative projects of self-knowledge, group knowledge, institutional knowledge – whose ends are not the tracking of data for its own sake, or even for profit, but for broader social, civic, cultural or political goals' (Couldry, 2014: 892; see also http://storycircle.co.uk/resources/social-analytics/). Social organisations' uses of analytics for these 'social, civic, cultural or political' purposes might be seen as a concrete example of the democratisation not just of data but of data mining.

Efforts to make social data open in the same ways as other types of data raise issues of ethics and personal privacy, discussed above, something which again underscores the need for public debate about and regulation of data mining. As Baack (2015) points out, open data advocates believe that personal data (that is, data which allows persons to be identified) should not be made open in the same way as other data. Nonetheless, some researchers have experimented with opening up social media data, such as Pybus et al. (forthcoming), who discuss their project 'Our Data, Ourselves' in this special issue. This project aimed to confront questions of agency in relation to social media data, given what the researchers saw as asymmetrical power relationships with regard to who gets to own the social data that we are all active in producing, and to explore how gaining access to one's own social data might augment agency. Aware that accessibility issues relate not just to data but also to software and the technical skills and expertise required to analyse data (see also Boyd and Crawford, 2012, discussed above), they worked with already-skilled young coders to create apps to intervene in social data produced and mined on mobile phones. Examples of apps created by the young coders include one designed to highlight the frequency of data tracking through audio alerts and another which produced graphs demonstrating the relationship between social media platform usage and frequency of data mining. At the time of writing, the project is only recently finished, so the results and findings have not yet been widely shared, but it nonetheless represents a concrete example of how social media data and their mining might be made more accessible to publics.

The third way in which data mining might be democratised relates to the types of publics it produces and the implications this has for the quality of the public sphere (Barnett, 2003: 54–80, 2008; Habermas, 1997: 329–387) or 'the mediapolis' (Silverstone, 2007). As noted earlier, the representations of publics generated through data mining circulate in the public sphere and constitute publics by shaping how publics come to be viewed and understood (Anstead and O'Loughlin, 2015; Gillespie, 2014). In one respect, this is not new: several democratic and media theorists have noted how 'the public' and 'public opinion' are not things that can be known without various technologies and practices of representation that make them present (Barnett, 2008; Osborne and Rose, 1999; Peters, 1995). As Barnett (2008: 404) puts it, 'people speak about what "the public" thinks, feels, and favours, and when they do so, they tend to have recourse to the results of elections, or statistical surveys, or opinion polls. These technical mediums are the ways in which the voice of the public is often expressed'.

However, while publics depend upon various forms of representation, not all ways of representing publics are equal and the implications different representations have for the quality of the public sphere vary significantly. In his account of 'the mediapolis', Silverstone (2007: 29) stresses how media representations exclude as well as include and stifle as well as promote public understanding and debate. Yet, he argues that good media are a crucial source of public reflexivity. Media, he writes, 'are not only the locus of reflexivity in this, the late modern world, but they are one of its key stimuli, and they themselves, at best, provide the materials for that reflection and criticism' (Silverstone, 2007: 20). He therefore urges us to consider the 'faculties of judgement and imagination, and the capacity of the mediapolis to provide, and enable, the resources for the exercise of both in the pursuit of more effective understanding and participation in the world' (Silverstone, 2007: 43).

Data mining and analytics (and the representations and visualisations of publics they generate) could provide an invaluable cognitive resource for members of the public to understand each other, reflect on matters of shared concern, and to decide how to act together as publics. In this way, data mining and analytics could help to provide 'shared references points' for publics that cut across the data-driven personalisation of digital space (Couldry and Turow, 2014). However, data mining is more often viewed as a way of generating knowledge about publics rather than for or with publics. To date, most attention has been focused on how data mining can capture what publics say and do, rather than how – as part of 'the mediapolis' (Silverstone, 2007) or public sphere – it can help members of the public to understand public issues and each other better, such that more informed and knowing publics may take shape. But there are some exceptions. We point below to examples where data mining has been used to promote public understanding and debate in ways that might enable the creation of more active and knowing publics.

Good data journalism uses data mining (and, relatedly, data visualisation) to promote public understanding of public issues. One example is the work done by *The Guardian* newspaper, in collaboration with academic researchers, on the Reading the Riots project, which analysed data about the riots that took place in the UK in the summer of 2011 (Vis, 2012). The project aimed to identify why looting took place, in response to the absence of a government enquiry into the causes of the riots. Combining data about the location of riots with deprivation data allowed the project team to question the dominant narrative which asserted that there was no link between the riots and poverty. A parallel analysis of the role social media played in the riots found that Twitter was not used to organise people to go looting, as was widely reported, but rather to respond to the riots: #riotcleanup, used to mobilise people to clean up the streets after the riots, was one of the most popular hashtags during the period (Vis, 2012). In this example, sharing mined social media data with the public and presenting alternative representations to those that dominated debate could be seen as an effort to increase publics' understandings of themselves as publics, and so to facilitate greater reflexivity among publics.

The analytics practices undertaken by social actors on the Storycircle project, discussed above, might also be seen as efforts to produce reflective and therefore knowing publics, in that data mining for 'social, civic, cultural or political goals' might mean that social groups come to know themselves in more reflexive ways. Other social researchers have also reflected on how digital methods like social media data mining may be used to involve publics in the process of knowledge construction. This possibility leads Marres (2012: 141) to argue that 'digitization may be unsettling established divisions of labour in social research'. She writes that:

> As online social research forces us to acknowledge the contributions of digital devices, practices and subjects, to the enactment of social research, it can be taken as an invitation to move beyond ''proprietary'' concepts of methods, that is, beyond the entrenched use of method as a way to monopolise the representation of a given field or aspect of social reality. (Marres, 2012: 160–161)

Likewise, Housley et al. (2014: 4) have argued that digital methods provide social researchers with new ways of collaborating with publics, making both a 'public sociology' and 'citizen social science' possible. Collaborative Online Social Media Observatory (COSMOS), an academic project undertaken by these authors, aims to operate as a 'collaboratory' where publics and researchers aim to produce knowledge together. As Housley et al. (2014: 12) describe it, the project's aim 'is to develop the COSMOS platform as a ''collaboratory'', an element of participatory research infrastructure supporting public engagement in a range of activities that includes the exchange of ideas, debates about the shape of institutions, current social problems, opportunities and events, as well as the co-production of social scientific knowledge through citizen social science, where publics act as vital sensors and interpreters of social life'.

These examples highlight embryonic practices which seek to use data mining in a way that contributes more positively to 'the mediapolis' (Silverstone, 2007). We noted above how Gillespie (2014: 189) points to a possible 'friction between ''networked publics'' forged by users and the ''calculated publics'' offered by algorithms'. But if data mining is used to enable the creation of more reflexive and active publics in these ways, the relationship between 'networked publics' and 'calculated publics' may be understood in more productive terms, in so far as knowledge generated through data mining is drawn upon by 'networked publics' in order to understand themselves, each other and public issues. As Peters (1995: 16) suggests, 'in acting upon symbolic representations of ''the public'' the public can come into existence as a real actor'. Understood this way, data mining and analytics are not only mechanisms for knowing publics, but can be means by which publics can know themselves.

Conclusion: Social media data mining and public agency

In this paper, we have considered possible responses to some of the mains criticisms that have been levelled at contemporary forms of data mining. In doing so, we have attempted to move beyond critique by considering whether a better relationship between social media data mining and public life might be possible and what alternative arrangements in relation to social media data mining might look like. We have argued that this requires a move from 'known publics', who are subject to the data-mining practices of others, to 'knowing publics', who are positioned in relation to data as more active and reflexive agents.

Concerns about data mining (in terms of increased surveillance and privacy invasion, related opaque forms of discrimination, social sorting and control over the way the public is represented) are now being followed by calls to do data mining differently and democratise data power. We have argued that there are three main ways in which this might be done. Firstly, we noted that commentators call for data mining to be transparent and open to public supervision and to be regulated by public authorities. Greater transparency and accountability of data mining are proposed as mechanisms to facilitate public understanding, debate and action in relation to data mining. Similarly, to address concerns about access to data practices, technologies and expertise, there have been calls for data and data mining to be more accessible and available as a common public good. Open data movements offer one (albeit not straightforward) example of this principle in action, but as yet, considerations of how it might apply to data *mining* and to *social media* data are somewhat limited. Finally, we suggested building on the notion that publics exist, in part, through the way they are represented, in order to consider whether data-mining practices can be used by publics to constitute themselves as more active and reflexive agents. Through these moves, it may be possible for data mining not only to be used by elites to produce known publics, but rather for the public to be more knowing of itself and to participate in the active production of itself, the public.

By reflecting on these issues, we have returned to the issue of agency, central to so many studies which have sought to explore how cultures and societies are made, and how they might be made fairer and more equal. In debates about which has primacy, structures or agency, structuralist critics would argue that structures not only determine, but serve to oppress and restrict the agency of already-disadvantaged groups in society. Some of the critics discussed in our paper fall into this category. In contrast, others have stressed the capacity of human agents to make and shape their worlds, albeit in the context of constraining structures. Others still have highlighted the dialectical relationship between structure and agency: structures shape and constrain human agency, but human agents act against, as well as within, them. We share this view. Cultural critic Jeremy Gilbert advocates such a position, a perspective which, he says, acknowledges 'the potency of both of these modes of analysis and the fact that they can both be true simultaneously'. In fact, he goes on to argue, 'I want to insist that we can't understand how capitalist culture works without understanding that they *are* both true' (Gilbert, 2012). It is within this ever-present, dialectic tension between structure and agency that this paper is situated. Returning to Couldry's (2014: 891) definition of agency as 'the longer processes of action based on reflection, giving an account of what one has done, even more basically, making sense of the world *so as* to act within it', we maintain that greater public agency in relation to data mining might be possible, under the conditions discussed in the second half of this paper. However far away we might currently be from realising these imaginings, it is certainly worth having them in sight.

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: All authors have agreed to the submission of this article to *Big Data and Society*. The article is not being considered for publication by any other print or electronic journal.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: the Engineering and Physical Sciences Research Council (EPSRC) under Grant 95575108.

References

Andrejevic M (2013) *Infoglut: How too Much Information is Changing the Way We Think and Know*. New York, NY: Routledge.

Andrejevic M (2011) The work that affective economics does. *Cultural Studies* 25(4–5): 604–620.

Andrejevic M and Gates K (2014) Big Data surveillance: Introduction. *Surveillance and Society* 12(2): 185–196.

Anonyzious (2012) 10 largest databases of the world. In: realitypod.com. Available at: http://realitypod.com/2012/03/10-largest-databases-of-the-world/ (accessed 16 October 2014).

Anstead N and O'Loughlin B (2012) Semantic polling: The ethics of online public opinion. Available at: http://eprints.lse.ac.uk/46944/1/LSEMPPBrief5.pdf (accessed 27 April 2013).

Anstead N and O'Loughlin B (2015) Social media analysis and public opinion: The 2010 UK general election. *Journal of Computer-Mediated Communication* 20(2): 204–220.

Baack S (2015) Datafication and empowerment: How the open data movement re-articulates notions of democracy, participation and journalism. *Big Data and Society*. DOI: 10.1177/2053951715594634.

Barnes T (2013) Big Data, little history. *Dialogues in Human Geography* 3(3): 297–302.

Barnett C (2003) *Culture and Democracy: Media, Space and Representation*. Edinburgh: Edinburgh University Press.

Barnett C (2008) Convening publics: The parasitical spaces of public action. In: Cox K, Low M and Robinson J (eds) *The Handbook of Political Geography*. London: Sage.

Barocas S, Hood S and Ziewitz M (2013) Governing algorithms: A provocation piece. Paper prepared for the "Governing Algorithms" conference, 16–17 May 2013, New York University. Available at: http://governingalgorithms.org/resources/provocation-piece/ (accessed 4 June 2013).

Barocas S and Selbst A (2014) Big Data's disparate impact. Social Science Research Network, SSRN. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id = 2477899 (accessed 2 October 2014).

Bates J (2013) Information policy and the crises of neoliberalism: The case of open government data in the UK. In: *Proceedings of the IAMCR 2013 conference*, Dublin, Ireland, 25–29 June. Available at: http://eprints.whiterose.ac.uk/77655/7/WRR0_77655.pdf (accessed 29 September 2014).

Beer D and Burrows R (2013) Popular culture, digital archives and the new social life of data. *Theory, Culture & Society* 30(4): 47–71.

Blanke T and Coté M (forthcoming) Hacking the social life of data: A data literacy framework. *Big Data and Society*.

Boyd D (2010) Social network sites as networked publics: Affordances, dynamics, and implications. In: Papacharissi Z (ed.) *Networked Self: Identity, Community and Culture on Social Network Sites*. New York, NY: Routledge.

Boyd D (2014) *It's Complicated: The Social Lives of Networked Teens*. New Haven, CT: Yale University Press.

Boyd D and Crawford K (2012) Critical questions for Big Data: Provocations for a cultural, technological and scholarly phenomenon. *Information, Communication and Society* 15(5): 662–679.

Couldry N (2014) Inaugural: A necessary disenchantment: Myth, agency and injustice in a digital world. *The Sociological Review* 62(4): 880–897.

Couldry N and Powell A (2014) Big Data from the bottom up. *Big Data & Society*. Epub ahead of print 2014. DOI: 10.1177/2053951714539277.

Couldry N and Turow J (2014) Advertising, Big Data and the clearance of the public realm: Marketers' new approaches to the content subsidy. *International Journal of Communication* 8: 1710–1726.

Diakopoulos N (2014) Algorithmic accountability reporting: On the investigation of black boxes. Tow Center for Digital Journalism. Available at: http://www.nickdiakopoulos.com/wp-content/uploads/2011/07/Algorithmic-Accountability-Reporting_final.pdf (accessed 6 October 2015).

Feenberg A (2002) *Transforming Technology: A Critical Theory Re-visited*. Oxford: Oxford University Press.

Freedman D (2012) Outsourcing internet regulation. In: Curran J, Fenton N and Freedman D (eds) *Misunderstanding the Internet*. London: Routledge.

Fuchs C (2014) *Social Media: A Critical Introduction*. London: Sage.

Giddens A (1984) *The Constitution of Society: Outline of the Theory of Structuration*. Oxford: Polity Press.

Gilbert J (2012) Moving on from market society: culture (and cultural studies) in a post-democratic age. *Open Democracy*. Available at: http://www.opendemocracy.net/ourkingdom/jeremy-gilbert/moving-on-from-market-society-culture-and-cultural-studies-in-post-democra (accessed 26 July 2012).

Gillespie T (2012) Can an algorithm be wrong? In: Limn 2. Available at: http://limn.it/can-an-algorithm-be-wrong/ (accessed 4 August 2014).

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ and Foot KA (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press. Available at: http://www.tarletongillespie.org/essays/Gillespie%20-%20The%20Relevance%20of%20Algorithms.pdf (accessed 12 January 2015).

Gurstein MB (2011) Open data: Empowering the empowered or effective data use for everyone? *First Monday* 16: 2–7. Available at: http://journals.uic.edu/ojs/index.php/fm/article/view/3316/2764 (accessed 20 January 2015).

Habermas J (1997) *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. London: Polity Press.

Hearn A (2010) Structuring feeling: Web 2.0, online ranking and rating, and the digital 'reputation' economy. *Ephemera: Theory & Politics in Organisation* 10(3/4) Available at: http://www.ephemeraweb.org/ (accessed 27 February 2011).

Hintz (2014) Policy hacking: hackathons and policy code, Communication for Empowerment: citizens, markets, innovations. In: *Proceedings of the ECREA Conference*, Lisbon, Portugal, 14 November.

Housley W, Procter R, Edwards A, et al. (2014) Big and broad social data and the sociological imagination: A collaborative response. *Big Data and Society*. Epub ahead of print 2014. DOI: 10.1177/2053951714545135

Johnson B (2010) Privacy no longer a social norm, says Facebook founder. *The Guardian*. Available at: http://www.theguardian.com/technology/2010/jan/11/facebook-privacy (accessed 7 May 2011).

Kant T (2014) Giving the 'viewser' a voice? Situating the individual in relation to personalization, narrowcasting, and public service broadcasting. *Journal of Broadcasting & Electronic Media* 58(3): 381–399.

MacKinnon R (2014) Companies need to work harder to keep the internet open. In: ft.com. Available at: http://www.ft.com/cms/s/0/e9883160-3a9f-11e4-bd08-00144feab dc0.html#axzz3lQSmXzLZ (accessed 15 September 2014).

Manovich L (2011) Trending: The promises and the challenges of big social data. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (Also in Gold MK (ed.) *Debates in the Digital Humanities*) (accessed 9 October 2013).

Marres N (2012) The redistribution of methods: On intervention in digital social research, broadly conceived. *The Sociological Review* 60: 139–165.

McQuail D (2003) *Media Accountability and Freedom of Publication*. Oxford: Oxford University Press.

Osborne T and Rose N (1999) Do the social sciences create phenomena? The example of public opinion research. *British Journal of Sociology* 50: 367–396.

Pariser E (2011) *The Filter Bubble: What the Internet is Hiding from You*. London: Penguin Press.

Pasquale F (2015) *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge, MA: Harvard University Press.

Peters JD (1995) Historical tensions in the concept of public opinion. In: Glasser TL and Salmon CT (eds) *Public Opinion and the Communication of Consent*. Guilford Press.

Rae A (2014) Open data visualization: The dawn of understanding? In: StatsLife Blog, Royal Statistical Society. Available at: http://www.statslife.org.uk/opinion/1815-open-data-visualisation-the-dawn-of-understanding (accessed 29 September 2014).

Raynes-Goldie K (2010) Aliases, creeping, and wall cleaning: Understanding privacy in the age of Facebook. *First Monday* 15(1). Available at: http://firstmonday.org/article/view/2775/2432 (accessed 19 June 2013).

Sandvig C, Hamilton K, Karahalio K, et al. (2014) Auditing algorithms: Research methods for detecting discrimination on internet platforms. Available at: http://www-personal.umich.edu/∼csandvig/research/Auditing%20Algorithms%20–%20Sandvig%20–%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (accessed 30 May 2014).

Silverstone R (2007) *Media and Morality: On the Rise of the Mediapolis*. Cambridge: Polity Press.

Trottier D (2012) *Social Media as Surveillance: Rethinking Visibility in a Converging World*. Farnham: Ashgate Press.

Turow J (2012) *The Daily You: How the New Advertising Industry is Defining Your Identity and Your Worth*. New Haven: Yale University Press.

Van Dijck J and Poell T (2013) Understanding social media logic. *Media and Communication* 1(1): 2–14.

Van Zoonen L (2014) Data delirium. *Sociologie Magazine* 22(3): 10.

Vis F (2012) *The Guardian* DataBlog's coverage of the UK riots. In: Gray L, Bounegru L and Chambers L (eds) *The Data Journalism Handbook Online*. Available at: http://datajournalismhandbook.org/1.0/en/index.html (accessed 12 November 2013).

Williamson B (2014) The death of the theorist and the emergence of data and algorithms in digital social research. Impact of Social Sciences. In: LSE Blog. Available at: http://blogs.lse.ac.uk/impactofsocialsciences/2014/02/10/the-death-of-the-theorist-in-digital-social-research/ (accessed 11 February 2014).

Young A and Quan-Haase A (2013) Privacy protection strategies on Facebook. *Information, Communication and Society* 16(4): 479–500.

Young IM (2002) *Inclusion and Democracy*. Oxford: Oxford University Press.

Zarsky TZ (2004) Desperately seeking solutions: Using implementation-based solutions for the troubles of information privacy in the age of data mining and the internet society. *Maine Law Review* 56: 13.

This article is part of a special theme on *Data and Agency*. To see a full list of all articles in this special theme, please click here: http://bds.sagepub.com/content/data-agency.